

MCGINN & GIBB, PLLC
A PROFESSIONAL LIMITED LIABILITY COMPANY
PATENTS, TRADEMARKS, COPYRIGHTS, AND INTELLECTUAL PROPERTY LAW
8321 OLD COURTHOUSE ROAD, SUITE 200
VIENNA, VIRGINIA 22182-3817
TELEPHONE (703) 761-4100
FACSIMILE (703) 761-2375; (703) 761-2376

**APPLICATION
FOR
UNITED STATES
LETTERS PATENT**

APPLICANT: **GUSTAVSON, ET AL.**

FOR: **METHOD AND STRUCTURE FOR
PRODUCING HIGH PERFORMANCE
LINEAR ALGEBRA ROUTINES USING
REGISTER BLOCK DATA FORMAT
ROUTINES**

DOCKET NO.: **YOR920030169US1**

METHOD AND STRUCTURE FOR PRODUCING HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING REGISTER BLOCK DATA FORMAT ROUTINES

DESCRIPTION

5 BACKGROUND OF THE INVENTION

Cross-Reference to Related Applications

The following seven Applications, including the present Application, are related:

1. U.S. Patent Application No. 10/ __, __, filed on ____, to
10 Gustavson et al., entitled "METHOD AND STRUCTURE FOR PRODUCING
HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING
COMPOSITE BLOCKING BASED ON L1 CACHE SIZE", having IBM Docket
YOR920030010US1;
2. U.S. Patent Application No. 10/ __, __, filed on ____, to
15 Gustavson et al., entitled "METHOD AND STRUCTURE FOR PRODUCING
HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING A HYBRID
FULL PACKED STORAGE FORMAT", having IBM Docket
YOR920030168US1;
3. U.S. Patent Application No. 10/ __, __, filed on ____, to
20 Gustavson et al., entitled "METHOD AND STRUCTURE FOR PRODUCING
HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING REGISTER
BLOCK DATA FORMAT", having IBM Docket YOR920030169US1;
YOR920030169US1

4. U.S. Patent Application No. 10/ ____, ____, filed on ____, to
Gustavson et al., entitled "METHOD AND STRUCTURE FOR PRODUCING
HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING LEVEL 3
PREFETCHING FOR KERNEL ROUTINES", having IBM Docket
5 YOR920030170US1;

5. U.S. Patent Application No. 10/ ____, ____, filed on ____, to
Gustavson et al., entitled "METHOD AND STRUCTURE FOR PRODUCING
HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING
PRELOADING OF FLOATING POINT REGISTERS", having IBM Docket
10 YOR920030171US1;

6. U.S. Patent Application No. 10/ ____, ____, filed on ____, to
Gustavson et al., entitled "METHOD AND STRUCTURE FOR PRODUCING
HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING A
SELECTABLE ONE OF SIX POSSIBLE LEVEL 3 L1 KERNEL ROUTINES",
15 having IBM Docket YOR920030330US1; and

7. U.S. Patent Application No. 10/ ____, ____, filed on ____, to
Gustavson et al., entitled "METHOD AND STRUCTURE FOR PRODUCING
HIGH PERFORMANCE LINEAR ALGEBRA ROUTINES USING
STREAMING", having IBM Docket YOR920030331US1, all assigned to the
20 present assignee, and all incorporated herein by reference.

YOR920030169US1

Field of the Invention

The present invention relates generally to techniques for improving performance for linear algebra routines. More specifically, a register block data format provides a method to mask a hardware-level instruction shortcoming such as a lack of a desired conventional hardware/assembly instruction to bring matrix data into Floating Point Registers in a desired matrix transpose format.

Description of the Related Art

Scientific computing relies heavily on linear algebra. In fact, the whole field of engineering and scientific computing takes advantage of linear algebra for computations. Linear algebra routines are also used in games and graphics rendering. Typically, these linear algebra routines reside in a math library of a computer system that utilizes one or more linear algebra routines as a part of its processing. Linear algebra is also heavily used in analytic methods that include applications such as a supply chain management.

A number of methods have been used to improve performance from new or existing computer architectures for linear algebra routines. However, because linear algebra permeates so many calculations and applications, a need continues to exist to optimize performance of matrix processing.

More specific to the technique of the present invention and as recognized by the present inventors, performance loss can occur for linear algebra processing

YOR920030169US1

in new computer architectures in which one or more conventional hardware or assembly instructions are lacking or deliberately excluded in order to reduce chip complexity and cost.

5

SUMMARY OF THE INVENTION

In view of the foregoing and other exemplary problems, drawbacks, and disadvantages of the conventional systems, it is, therefore, an exemplary feature of the present invention to provide various techniques that improve performance for linear algebra routines.

10

It is another exemplary feature of the present invention to provide a method and structure to provide higher-performance linear algebra routines for environments in which one or more conventional hardware or assembly instructions are missing.

15

It is another exemplary feature of the present invention to deal with new computer architectures and assembly languages in a better manner than previous methods and to demonstrate a general methodology for masking architectural shortcomings.

20

In a first exemplary aspect of the present invention, described herein is a computerized method of executing a matrix operation, including, for a matrix A, separating the matrix A into blocks, each block having a size p-by-q, and storing the matrix data in a cache or memory in a different format. The new format is at YOR920030169US1

least one of a new format in which elements of a block are stored in a different location relative to its original position in the block and/or a new format in which blocks of size p -by- q are stored such that at least one block occupies a different position relative to its original position in the matrix A .

5 In a second exemplary aspect of the present invention, described herein is an apparatus including a reader to read a data of a matrix A , a separator to separate the data into blocks of a size p -by- q , a calculator to calculate a new position of block elements and/or blocks that differ from an original position in the matrix A . A memory loader stores the blocks into a memory, the different
10 positioning of blocks being a register block data format of the matrix A .

 In a third exemplary aspect of the present invention, described herein is a data structure in a computer program executing a matrix operation, the data structure including, for a matrix A , separating the matrix A into blocks, and storing the data into a memory or cache according to a new format as described
15 above.

 In a fourth exemplary aspect of the present invention, described herein is a signal-bearing medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method of storing information of a matrix in accordance with the above-described method.

20 In a fifth exemplary aspect of the present invention, described herein is a

YOR920030169US1

method of providing a service that includes at least one of: using a linear algebra software package that computes one or more matrix subroutines, wherein the linear algebra software package processes a matrix data for a matrix A in accordance with the above-described method; providing a consultation for solving a scientific/engineering problem using the linear algebra software package; transmitting a result of the linear algebra software package on at least one of a network, a signal-bearing medium containing machine-readable data representing the result, and a printed version representing said result; and receiving a result of the linear algebra software package on at least one of a network, a signal-bearing medium containing machine-readable data representing the result, and a printed version representing the result.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other exemplary features, aspects and advantages will be better understood from the following detailed description of exemplary embodiments of the invention with reference to the drawings, in which:

Figure 1 illustrates an exemplary linear algebra operation 100;

Figure 2 illustrates an exemplary hardware/information handling system 200 for incorporating the present invention therein;

Figure 3 exemplarily illustrates a CPU 211 that includes a floating point unit (FPU) 302;

YOR920030169US1

Figure 4 exemplarily illustrates in more detail the CPU 211 that might be used in the computer system 200 for the present invention;

Figure 5 illustrates an exemplary pair of floating point registers 501, 502 in a floating point unit register bank;

5 Figure 6 illustrates pictorially the concept of converting a matrix A into the pseudo-matrix format 605;

Figure 7 illustrates a method 700 for the conversion of matrix A into pseudo-matrix format and the subsequent loading of the pseudo-matrix format into a series of data registers; and

10 Figure 8 illustrates a signal bearing medium 800 (e.g., storage medium) for storing steps of a program of a method according to the present invention.

DETAILED DESCRIPTION OF EXEMPLARY

EMBODIMENTS OF THE INVENTION

15 Referring now to the drawings, and more particularly to Figures 1-8, an exemplary embodiment will now be described. In this exemplary embodiment, matrix data will be stored in a modified format in memory, which the inventors choose to call an example of "register block data format". In the special case where the block size is 2-by-2, the register block data format is referred to as a

20 "pseudo-matrix". The matrix, partitioned into register block data format, is brought into a floating point unit (FPU) in which the floating point registers

YOR920030169US1

(FRegs) are viewed in a specific manner (e.g., "checkerboarding") so that a combination of two "errors" can be used to overcome lack of a hardware or assembly instruction for bringing the matrix into the FPU in a preferred format.

The present invention addresses, generally, efficiency in the calculations of linear algebra routines. Figure 1 illustrates processing of an exemplary matrix operation 100 (e.g., $C = C - A^T * B$). In processing this operation, matrix A is first transposed to form transpose-matrix-A (e.g., A^T) 101. Next, transposed matrix A^T is multiplied with matrix B 102 and then subtracted from matrix C 103. The computer program executing this matrix operation will achieve this operation using three loops 104 in which the element indices of the three matrices A, B, C will be varied in accordance with the desired operation.

That is, as shown in the lower section of Figure 1, the inner loop and one step of the middle loop will cause indices to vary so that MB rows 105 of matrix A^T will multiply with NB columns 106 of matrix B. The index of the outer loop will cause the result of the register block row/column multiplications to then be subtracted from the MB-by-NB submatrix 107 of C to form the new submatrix 107 of C. Figure 1 shows an exemplary "snapshot" during execution of one step of the middle loop $i = i: i + MB - 1$ and all steps of the inner loop l, with the outer loop $j = j: j + NB - 1$.

The above matrix operation $C = C - A^T * B$ is now examined in more depth. At any point the operation is: $C(i: i + MB - 1, j: j + NB - 1) = C(i: i + MB - 1,$
YOR920030169US1

j:j + NB-1) - A(l:l + KB-1, i:i + MB-1) * B(l:l + KB-1, j:j + MB-1). Using standard column major format for C, A^T , and B would mean that the elements in submatrices C(i:i + MB-1, j:j + NB -1), A^T (l:l + KB-1, i:i + MB-1) and B(l:l + KB -1, j:j + NB-1) are not stored contiguously in some order that is advantageous for the FPU.

Therefore, a key idea of the present invention is to store these three submatrices contiguously in some representation (permutation) that has optimal advantage for the L1 cache-FPU register interface of a particular architecture.

The C, A^T , B submatrices have $CE = MB * NB$, $AE = KB * MB$, and $BE = KB * NB$ elements, respectively. Thus, there are $CE!$, $AE!$, $BE!$ stride one representations to choose from in order to achieve an optimal advantage for the L1 cache-FPU register interface. In accordance with the present invention, for a given architecture, a particular MB, NB, KB will be chosen and then a particular permutation from each of the possible $CE!$, $AE!$, $BE!$ choices will be chosen.

It is also noted that the above-listed fourth and fifth copending applications on preloading and prefetching, as well as the sixth application on six choices of the kernel routines (in the present application, only one of the six possible kernels are focused on), are all related, but independent, aspects of obtaining optimal performance for matrix multiplication.

In the context of the present invention, the prefetching discussed elsewhere retrieves the C, A^T , B submatrix blocks to L1 ahead of time. The

YOR920030169US1

preloading technique transfers the C, A^T , B submatrix blocks to the FPU registers in a timely manner for the FPU executions and subsequent transfer of the FPU results back to the L1 cache, higher order caches, or the memory.

5 Finally, for the choice of having six kernels, instead of one, allows for having six choices to optimize. This optimization depends on a number of parameters, including, for example, whether the matrix processing program is oriented in row major or column major.

However, before presenting the exemplary details of the present invention, the following general discussion provides a background of linear algebra subroutines and computer architecture as related to the terminology used herein.

10

Linear Algebra Subroutines

The explanation of the present invention includes reference to the computing standard called LAPACK (Linear Algebra PACKage) and to various subroutines contained therein. Information on LAPACK is readily available on the Internet.

15

For purpose of discussion only, Level 3 BLAS (Basic Linear Algebra Subprograms) of the LAPACK (Linear Algebra PACKage) are used, but it is intended to be understood that the concepts discussed herein are easily extended to other linear algebra mathematical standards and math library modules.

20

When LAPACK is executed, the Basic Linear Algebra Subprograms (BLAS), unique for each computer architecture and provided by the computer

YOR920030169US1

vendor, are invoked. LAPACK comprises a number of factorization algorithms for linear algebra processing.

For example, Dense Linear Algebra Factorization Algorithms (DLAFAs) include matrix multiply subroutine calls, such as Double-precision Generalized Matrix Multiply (DGEMM). At the core of level 3 Basic Linear Algebra Subprograms (BLAS) are “L1 kernel” routines which are constructed to operate at near the peak rate of the machine when all data operands are streamed through or reside in the L1 cache.

The most heavily used type of level 3 L1 DGEMM kernel is Double-precision A Transpose multiplied by B (DATB), that is, $C = C - A^T * B$, where A, B, and C are generic matrices or submatrices, and the symbology A^T means the transpose of matrix A. It is noted that DATB is the only such kernel employed by today’s state of the art codes, although DATB is only one of six possible kernels.

The DATB kernel operates so as to keep the A operand matrix or submatrix resident in the L1 cache. Since A is transposed in this kernel, its dimensions are $K1$ by $M1$, where $K1 \times M1$ is roughly equal to the size of the L1. Matrix A can be viewed as being stored by row, since in Fortran, a non-transposed matrix is stored in column-major order and a transposed matrix is equivalent to a matrix stored in row-major order. Because of asymmetry (C is both read and written) $K1$ is usually made to be greater than $M1$, as this choice leads to superior performance.

YOR920030169US1

Exemplary Computer Architecture

Figure 2 shows a typical hardware configuration of an information handling/ computer system 200 usable with the present invention. Computer system 200 preferably includes at least one processor or central processing unit (CPU) 211. Any number of variations are possible for computer system 200, including various parallel processing architectures and architectures that incorporate one or more FPUs (floating-point units).

In the exemplary architecture of Figure 2, the CPUs 211 are interconnected via a system bus 212 to a random access memory (RAM) 214, read-only memory (ROM) 216, input/output (I/O) adapter 218 (for connecting peripheral devices such as disk units 221 and tape drives 240 to the bus 212), user interface adapter 222 (for connecting a keyboard 224, mouse 226, speaker 228, microphone 232, and/or other user interface device to the bus 212), a communication adapter 234 for connecting an information handling system to a data processing network, the Internet, an Intranet, a personal area network (PAN), etc., and a display adapter 236 for connecting the bus 212 to a display device 238 and/or printer 239 (e.g., a digital printer or the like).

Although not specifically shown in Figure 2, the CPU of the exemplary computer system could typically also include one or more floating-point units (FPUs) that performs floating-point calculations. Computers equipped with an FPU perform certain types of applications much faster than computers that lack

YOR920030169US1

one. For example, graphics applications are much faster with an FPU. An FPU might be a part of a CPU or might be located on a separate chip. Typical operations are floating point arithmetic, such as fused multiply/add (FMA), addition, subtraction, multiplication, division, square roots, etc.

5 Details of the FPU is not so important for an understanding of the present invention, since a number of configurations are well known in the art. Figure 3 shows an exemplary typical CPU 211 that includes at least one FPU 302. The FPU function of CPU 211 controls the FMAs (floating-point multiply/add), and at least one load/store unit (LSU) 301, which loads/stores data to/from memory
10 device 304 into the floating point registers (FReg's) 303).

 It is noted that, in the pretext of the present invention involving linear algebra processing, the term "FMA" can also be translated either as "fused multiply-add" operation/unit or as "floating-point multiply followed by floating point add" operation/unit, and it is not important for the present discussion which
15 translation is used. The role of the LSU 301 is to move data from a memory device 304 external to the CPU 211 to the FRegs 303 and to subsequently transfer the results of the FMAs back into memory device 304, typically via the L1 cache. It is important to recognize that the LSU function of loading/storing data into and out of the FRegs 303 occurs in parallel with the FMA function.

20 Another important aspect of the present invention relates to computer architecture that incorporates a memory hierarchy involving one or more cache
YOR920030169US1

memories. Figure 4 shows in more detail how the computer system 200 might incorporate a cache 401 in the CPU 211.

Discussion of the present invention includes reference to levels of cache, and more specifically, level 1 cache (L1 cache). Level 1 cache is typically
5 considered as being a cache that is closest to the CPU and might even be included as a component of the CPU, as shown in Figure 4.

The details of the cache structure and the precise location of the cache levels are not so important to the present invention so much as recognizing that memory is hierarchical in nature in modern computer architectures and that matrix
10 computation can be enhanced considerably by modifying the storage representation of a matrix that is the data of matrix subroutines to include considerations of the memory hierarchy.

Additionally, in the present invention, it is preferable that the matrix data be laid out contiguously in memory in "stride one" form. "Stride one" means that
15 the data is preferably contiguously arranged in memory to honor double-word boundaries and that the useable data is retrieved in increments of the line size. This contiguous arrangement of data facilitates bringing the data from higher levels of memory or higher levels of cache into L1 cache.

However, in the present invention, additionally it is necessary to transfer
20 the L1 matrix data into FPU registers in an optimal manner.

YOR920030169US1

Various of the above-listed copending application address the availability of six kernel types in matrix processing. The importance of having six kernel types available is that stride one memory access is desirable for matrix processing. The matrices A and B are stored either by row or by column. By having six
5 kernel types, one can choose a kernel in which stride one is available for both operands.

In the next paragraphs on pseudo-transpose, it is shown how to re-order the $MB = 2$ and $KB = 2$ subblocks of A^T into one of $4! = 24$ permutations. The permutation selected, which is optimal for the given architecture, does not lead to
10 the standard row or column major representation used by programming languages Fortran and C. In terms of these languages, the arrangement is no longer a standard submatrix. Hence, a term used in the present invention is "pseudo".

Pseudo-Transpose

The pseudo-transpose data format of the present invention can be used to
15 mask a hardware-level instruction shortcoming (i.e., the lack of a desired hardware/assembly instruction).

As an example of a hardware "deficiency" to be overcome in software by the concept by the present invention is shown in Figure 5 and Figure 6. A word length is assumed exemplarily as being 32 bits (four bytes) and it is assumed that

YOR920030169US1

the application requires that data be processed in double-word (e.g., 64 bits) increments 501.

A CPU is typically designed around this concept of 32-bit words and would have 32-bit-wide registers. However, a floating-point processor unit (FPU) typically is designed to execute in double-word (64-bit) increments as the basic data processing unit and sometimes processes floating-point arithmetic calculations using two such double-word data units. That is, the FPU typically has an FReg 501, 502 register-width of 64 bits (i.e., double-word registers) and the LSU 301 retrieves/stores data between memory/registers in increments of 64 or 128 bits.

Moreover, in the design of the FPU 302 for an exemplary computer supported by the present invention, the FPU 64-bit register loading/storing is executed using a “quad-load” format, meaning that the four words to be loaded-into/stored-from two double-word registers 501, 502 are actually accessible in units of four single-word lengths. Thus, quad-loading causes two consecutive single-words to be loaded into FReg₀ 501 as Word₀ and Word₁ and two consecutive single-words to be loaded into FReg₁ 502 as Word₂ and Word₃. However, quad-loading also allows the double words to be flipped (reversed) so that, for example, the contents of Word₀/Word₁ and Word₂/Word₃ would be reversed.

YOR920030169US1

As shown in Figure 1, matrix A would typically be processed in a LAPACK linear algebra subroutine as A transpose (i.e., A^T), shown in Figure 6 as label 602. The problem encountered in the hardware of a computer-under-design, which resulted in an exemplary solution of the present invention, is that
5 kernel-level matrix processing of LAPACK would be typically based on submatrices that are 2-by-2 data blocks.

That is, a BLAS kernel would typically work with increments of 2-by-2 data blocks, such as illustrated by the dashed box 603. Unfortunately, conventional computer architecture/assembly instructions would be based on the
10 concept of memory transfers based on line size, more appropriately shown as dashed box 604, if matrix A 601 was stored in memory in stride-one format.

To overcome the deficiency that the computer-under-design had no special hardware/assembly instruction to break 2-by-2 blocks out of memory for loading into the FRegs, the present invention utilizes a "pseudo-matrix" concept shown in
15 Figure 6 as pseudo-matrix 605.

In essence, in the present invention, matrix A (601) is stored in memory in pseudo-matrix format 605, rather than transpose matrix format 602. It can be seen that the pseudo-matrix format 605 is based on considering 2-by-2 blocks as a basic pseudo element block. It should also be apparent from comparing normal
20 matrix A (601) with pseudo-matrix 605 that the 2-by-2 block 603 is actually in a "transposed" position relative to its adjacent 2-by-2 block 606.

YOR920030169US1

That is, for the 2-by-2 block 603, in the pseudo-matrix 605, this block 603 occupies the column major location from its original position in the matrix. It is noted that blocks on the diagonal (e.g., 2-by-2 block 606) will continue to occupy the original diagonal location in the pseudo-matrix 605.

5 Further, in comparing the pseudo-matrix 605 with the transposed matrix 602, required by the exemplary matrix processing algorithm in Figure 1, it can be seen that the transpose matrix 602 would now result from the pseudo-matrix 605 if the single words of each of the 2-by-2 blocks 603, 606 are loaded into specific predetermined words of the appropriate FRegs, and that the appropriate relocation
10 involves a crisscrossing of one of the diagonals.

That is, for example, in 2-by-2 block 606 in pseudo-matrix 605, it can be seen that the required transpose would result when diagonal elements "2" and "5" are crisscrossed, and in 2-by-2 block 603, the transpose occurs when diagonal elements "10" and "13" are crisscrossed.

15 In the present invention, this predetermined placement in the FRegs is accomplished by considering the words of the FRegs as a "checkerboard", in which quad-loading of the pseudo-matrix data is appropriately matched to words in the FRegs to thereby result in the transpose format 602 in the FRegs. Therefore, even though no special computer instruction exists in the computer
20 instruction set that will allow a matrix transpose to be loaded into the FRegs efficiently, by first loading matrix A into the novel format of the present
YOR920030169US1

invention, followed by a modified version of quad-loading, the matrix A can be efficiently loaded into FRegs in the required transpose format.

Thus, in effect, the present invention makes two "errors" to arrive at loading the matrix A into transpose matrix format 602 in the registers of the FPU.

5 First, the data is stored in memory in pseudo-matrix format 605. Second, the pseudo-matrix data 605 is loaded from memory into the FRegs using quad-loading, but the quad-loading is modified in a predetermined way so that the transpose format 602 results from the loading.

10 It should be apparent that the reverse could also be accomplished to store data back into memory from the FRegs.

It should also be apparent from the explanation for 2-by-2 blocks typically used for matrix processing, that the concept could be generalized. The generalization of "pseudo-transpose data format" is "register-block data format" and will be further discussed later.

15 Figure 7 shows an exemplary flowchart 700 of the pictorial conversion shown in Figure 6. In step 701, matrix A is separated into blocks of a predetermined size p-by-q (e.g., 2-by-2). As discussed above, for the exemplary repositioning of blocks that illustrate the present invention, in step 702, the matrix A is stored in register block format by storing the blocks in normal/transposed positions relative to the position of the register blocks of matrix A. This normal/transposed configuration is appropriate for solving the specific computer
20 YOR920030169US1

architectural problem presented to being solved by the present invention.

However, it should be apparent that any number of other repositioning configurations might be appropriate. As possible examples, repositioning might be based on a concept of diagonals or wrap-around. Steps 701 and 702 would typically be done simultaneously.

In step 703, the register block data is loaded into a register pair of data registers, using either a normal or a checkerboard quad load so as to result in having data in the register pair of data registers as being either the normal or the transpose form of the register block of matrix A.

In step 704, the contents of the register pair of data registers are used in a linear algebra operation, such as a LAPACK BLAS subroutine, with the result being stored in another pair of data registers or back into L1 cache.

The exemplary description above is now formalized in the case $p = q = 2$.

Definition: Pseudo-Matrix for SIMD (Single Instruction, Multiple Data stream processing) FPU case:

Let A^T consist of two-by-two submatrices, where element $A^T(i:i+1, l:l+1) = A(l:l+1, i:i+1)$. Now let each two-by-two block of PA hold $A^T(i:i+1, l:l+1)^T$. (See Figure 6, where A has order $n = 4$.)

Although this formal idea will be described in the context of a single processor machine, extensions to both SMP (Symmetric MultiProcessing) and uni-processor SIMD architectures are quite straightforward.

YOR920030169US1

For purposes of concreteness, an architecture with 32 register-pairs and a dual SIMD configuration will be considered. The design details of the level 3 (BLAS), L1 cache kernel routine are delineated (below, in computer code) for a single-processor machine and the design can be extended for this example. Thus,
5 it suffices to consider the level 3 L1 DGEMM kernel, as demonstrated below in code.

The pertinent fact is that one wishes to use a matrix multiplication kernel that uses mn registers for C , km registers for A , and kn registers for B , where $mn + k(m+n) < \#$ available machine registers. Further, it is wise to prepare the data so
10 that it can be accessed (streamed) through the L1 cache, as each element is used multiple times by the FPU.

The pseudo-transpose idea allows for the following extensions:

a) To machines with multiple SIMD FPUs, where the hardware/assembly-level instructions would preclude a straightforward extension
15 of this approach (i.e. from $k = 1$ to $k = 2$; also, $k_0 = 1$ to $k_0 = 2$ in the RMP function code section below);

b) Without paying any noticeable performance penalty (data preparation overhead is almost inconsequential, as it is combined with data preparation for streaming); and

20 c) With little to no additional coding or code maintenance work.

YOR920030169US1

For additional explanation, the Algorithms and Architectures approach is used to consider an example that arose in practice. It should be pointed out that the problem encountered here (in this example) can be seen in the hardware-language instruction sets for other, new architectures. From the algorithm side, 2-by-2 submatrices will be used to express the 4-by-4 by 1 DATB (double precision $A^T * B$) kernel in the Matlab code below:

L1 Kernel for Computing $C = A^T * B + C$:

```

function Cout = DATB( m1, n2, k1, AR, BC, CC )
    % L1 kernel for computing  $C = A^T * B + C$ 
    % m1, n2, and k1 are integer multiples of m0, n0, and k0.
    % CReg, AReg, and BReg represent subsets of the registers
    % used for C, A, and B, respectively.
    % Parameter naming note:
    % The second character indicates the major-ordering of the
    % matrices. Thus, AC is A in column-major ordering. In both
    % MATLAB and Fortran, column-major corresponds to non-
    % transpose, while row-major ordering reflects a transposed
    % state; i.e.  $A^T$  (or  $A'$  in MATLAB notation).
    for j=1:n0:n2
        for i=1:m0:m1
            CReg = CC( i:i+m0-1, j:j+n0-1 );    % load submatrix of C
            % Perform operation  $CReg = CReg + A^T(*,i:i+m0-1)*B(*,j:j+n0-1)$ 
            for p=1:k0:k1
                AReg = AR( p:p+k0-1, i:i+m0-1 );    % load panel of A
                BReg = BC( p:p+k0-1, j:j+n0-1 );    % load panel of B
                CReg = AReg' * BReg + CReg;          % perform operation
            End
            CC( i:i+m0-1, j:j+n0-1 ) = CReg;        % store submatrix of C
        end
    end
    Cout = CC

```

On the Hypothetical Machine (HM), this will require unrolling the inner (p) loop by 2, so that a 4-by-4-by-2 L1 DATB kernel is produced. The idea is extensible to other values of unrolling by logical extension.

YOR920030169US1

It is noted that "loop unrolling" refers to a function call replaced with an instance of the function's body. Actual arguments are substituted for formal parameters. An unroll function is usually done as a compile-time transformation to increase the efficiency of the program. That is, loop unrolling includes

5 replicating the body of a loop in order to reduce the number of iterations required to complete the loop and allows keeping operands in FRegs 303.

This can then be implemented by performing 2-by-2-by-1 submatrix computations on 2-by-2 submatrices. It is necessary to balance the use of the P(rietary) and S(econdary) register files. One way to do this is to use "red-black" or "checkerboard" register assignment. That is, $a(i:i+1,j)$ is assigned to (P,S)/(S,P)

10 if and only if $i+j = (0/1) \bmod 2$. It is noted that registers "P" and "S" correspond to "FReg₀" 501 and "FReg₁" 502 in Figure 5.

From the architecture side, it is desired to preferably use the HM as a SIMD vector machine, for vectors of length two. The A matrix is chosen to hold

15 vectors of length 2. The B matrix will also hold vectors of length 2. However, the components of B will be used as scale factors by this dual SIMD FPU. A simple example will clarify what is intended, as follows.

The first column of $C(i:i+1,j:j+1) = A(i:i+1,l:l+1) * B(l:l+1, j:j+1)$ is computed as:

20

$$\begin{matrix} a(i,l) \\ a(i+1,l) \end{matrix} * b(l,j) + \begin{matrix} a(i,l+1) \\ a(i+1,l+1) \end{matrix} * b(l+1,j)$$

YOR920030169US1

and the second column is calculated in a similar fashion, using the two scalars, $b(l:l+1, j+1)$. If $i+l = 0 \bmod 2$ then $a(i:i+1, l)$ is placed in a (P,S) register pair. Since $i+l+1 = 1 \bmod 2$ in this case, $a(i:i+1, l+1)$ is placed in an (S,P) register pair.

5 Normally, to compute $A*B$ quickly, a cache-resident A is necessary that "maps well" into L1 cache. Thus, A is stored preferably contiguously and is assumed accessible stride-one. Hence, it would be expected to transpose A preferably into a contiguous buffer in a data preparation phase. When this is done however, a real problem is encountered that appears to have no solution.

10 If a 2-by-2 submatrix, $A(l:l+1, i:i+1)$ of A^T is considered, the HM processor's load unit only allows loading consecutive storage elements as vectors. For example, $a(l:l+1, i)$ can be loaded into a (P,S) (or (S,P)) register pair (if column-major storage is assumed). However, to use the HM most efficiently, it is necessary to load, for example, $a(l, i)$ and $a(l, i+1)$ into a (P,S) register pair and $a(l+1, i)$ and $a(l+1, i+1)$ into an (S,P) register pair. This corresponds to the
15 aforementioned vector computation, with i and l interchanged because of transposition.

These two disparate requirements can be reconciled in various ways. For example, "bending the algorithm" could be performed to "accommodate the architecture." The idea is to not form A^T in the contiguous buffer as would be
20 done in the Matlab coding above. Rather, a pseudo matrix as mentioned earlier,

YOR920030169US1

which in reality is not a matrix, is formed. Also, as noted previously, this pseudo matrix is a special case of the general p-by-q register block matrix.

Let A^T include 2-by-2 submatrices, where element $A^T(i:i+1,l:l+1) = A(l:l+1,i:i+1)$. Now let each 2-by-2 block of PA hold $A^T(l,i+1,l:l+1)^T$. It should be clear that this "mistake" in computing A^T allows one to load the HM registers both correctly and most efficiently, as is described above. In effect, an example has been produced where "two wrongs" make a "right."

Next, a 2-by-2 by 2 kernel is illustrated below. The invention is computing "C = C - A*B". The actual Matlab code is given below.

```

10 Do J = 0, N-2, 2
    Do I = 0, M-2, 2
        P0, S0 = C(I:I+1,J)      ! P quad load
        S1, P1 = C(I:I+1,J+1)    ! S quad load
        Do L = 0, K-2, 2
            P8, S8 = B(L:L+1,J)    ! P Quad Load
            PC, SC = PA(L:L+1,I)    ! P Quad Load
            (P0,S0) <- (P0,S0) - (PC,SC) * P8 ! 00C8-11-01 DFMA
            S9, P9 = B(L:L+1,J+1)  ! S Quad Load
            (P1,S1) <- (P1,S1) - (SC,PC) * S9 ! 11C9-00-10 DFMA
            SD, PD = PA(L:L+1,I+1)  ! S Quad Load
            (P0,S0) <- (P0,S0) - (SD,PD) * S8 ! 00D8-00-10 DFMA
            (P1,S1) <- (P1,S1) - (PD,SD) * P9 ! 11D9-11-01 DFMA
        EndDo
        C(I:I+1,J) = P0, S0        ! P Quad Store
        C(I:I+1,J+1) = S1, P1     ! S Quad Store
    EndDo
EndDo

```

Note: A single HM operation has the syntax (D, B, A, C, BPS, BPA, BPC, BSS, BSA, BSC), where D,B,A,C are the four register numbers of a FPU (P,S) register pair and the remaining six entries are Boolean variables signifying sign and MUX settings of the (P,S), A, and C operands.

YOR920030169US1

The full 4-by-4 by 2 kernel that computes $C = C - A * B$ is made up of four 2-by-2 by 2 code fragments. The following register layouts are recommended:

5	$C(i:i+3, j:j+3) =$	<table border="0"> <tr><td>P0</td><td>S1</td><td>P4</td><td>S5</td></tr> <tr><td>S0</td><td>P1</td><td>S5</td><td>P5</td></tr> <tr><td>P2</td><td>S3</td><td>P6</td><td>S7</td></tr> <tr><td>S2</td><td>P3</td><td>S6</td><td>P7</td></tr> </table>	P0	S1	P4	S5	S0	P1	S5	P5	P2	S3	P6	S7	S2	P3	S6	P7
P0	S1	P4	S5															
S0	P1	S5	P5															
P2	S3	P6	S7															
S2	P3	S6	P7															
	$B(1:1+1, j:j+3) =$	<table border="0"> <tr><td>P8</td><td>S9</td><td>PA</td><td>SB</td></tr> <tr><td>S8</td><td>P9</td><td>SA</td><td>PB</td></tr> </table>	P8	S9	PA	SB	S8	P9	SA	PB								
P8	S9	PA	SB															
S8	P9	SA	PB															
10	$PA(1:1+1, i:i+3) =$	<table border="0"> <tr><td>PC</td><td>SD</td><td>PE</td><td>SF</td></tr> <tr><td>SC</td><td>PD</td><td>SE</td><td>PF</td></tr> </table>	PC	SD	PE	SF	SC	PD	SE	PF								
PC	SD	PE	SF															
SC	PD	SE	PF															

The outer loops on J and I would be unrolled by four. Before entering the inner loop, register pairs 0 to 7 would be loaded with 16 elements of C. The inner loop on L would be unrolled by two. Each pass through the inner loop would consist of two sets of two quad-loads of PA and four quad loads of B. These eight quad loads will feed 16 dual FMA (DMFA) instructions.

The resulting code, which is a simple extension of the single-processor code (this is true in both Matlab and C/Fortran) achieves laudable performance on real architectures. This idea will result in similar performance for other new architectures as well.

As a second example in which an architecture/assembly instruction would benefit from the present invention, the Intel P3 and P4 processors have SSE(1,2) instructions. These are Intel Steaming InStruction Extensions and allow for the execution of two floating point multipliers and two floating point additions (double precision) in a single cycle. Thus, use of the SSE# instructions provide a

YOR920030169US1

potential for a factor-of-two performance gain on floating point intensive computation. By using the pseudo transpose data format described above, the benefit of this 100% performance gain can be realized.

The pseudo-matrix discussed above is a special case of the register block data format in that its size is 2-by-2. That is, the generalization of "pseudo-matrix data format" is "register block data format", which is a p-by-q block where p and q are small integers so that the pieces of these such blocks can then be fitted into the registers of a particular architecture to achieve a desirable data format stored in these registers.

The layout of these blocks could be arbitrary. As noted, there are AE! choices for the submatrix block of A. As discussed above, in usual cases, the p-by-q sub-blocks will be laid out either in row- or column-major format. But a key idea of the present invention is that the arbitrary layout of these blocks is tailored to the architectural design of the FPU and its associated floating point registers FReg's. It should be apparent that different architectural or instruction set scenarios would provide a need to layout the blocks differently from the exemplary layout of the present invention. It is intended that these different layouts, as required by unique architectural/instruction-set combinations, be considered within the scope of the present invention, since the present invention is intended to teach that dividing matrix data into blocks and shifting locations of

YOR920030169US1

these blocks can be a basic technique to overcome "deficiencies" in computer architectural or instruction set design.

In addition to the hardware/software environment described above, a different exemplary aspect of the invention includes a computer-implemented method for performing the invention.

Such a method may be implemented, for example, by operating a computer, as embodied by a digital data processing apparatus, to execute a sequence of machine-readable instructions. These instructions may reside in various types of signal-bearing media.

Thus, this exemplary aspect of the present invention is directed to a programmed product, comprising signal-bearing media tangibly embodying a program of machine-readable instructions executable by a digital data processor incorporating the CPU 211 and hardware above, to perform the method of the invention.

This signal-bearing media may include, for example, a RAM contained within the CPU 211, as represented by the fast-access storage for example. Alternatively, the instructions may be contained in another signal-bearing media, such as a magnetic data storage diskette 800 (Figure 8), directly or indirectly accessible by the CPU 211.

Whether contained in the diskette 800, the computer/CPU 211, or elsewhere, the instructions may be stored on a variety of machine-readable data

YOR920030169US1

storage media, such as DASD storage (e.g., a conventional "hard drive" or a RAID array), magnetic tape, electronic read-only memory (e.g., ROM, EPROM, or EEPROM), an optical storage device (e.g. CD-ROM, WORM, DVD, digital optical tape, etc.), paper "punch" cards, or other suitable signal-bearing media
5 including transmission media such as digital and analog and communication links and wireless.

The second exemplary aspect of the present invention can be embodied in a number of variations, as will be obvious once the present invention is understood. That is, the methods of the present invention could be embodied as a
10 computerized tool stored on diskette 800 that contains a series of matrix subroutines to solve scientific and engineering problems using matrix processing in accordance with the present invention. Alternatively, diskette 800 could contain a series of subroutines that allow an existing tool stored elsewhere (e.g., on a CD-ROM) to be modified to incorporate one or more of the principles of the
15 present invention.

The second exemplary aspect of the present invention additionally raises the issue of general implementation of the present invention in a variety of ways.

For example, it should be apparent, after having read the discussion above that the present invention could be implemented by custom designing a computer
20 in accordance with the principles of the present invention. For example, an

YOR920030169US1

operating system could be implemented in which linear algebra processing is executed using the principles of the present invention.

In a variation, the present invention could be implemented by modifying standard matrix processing modules, such as described by LAPACK, so as to be based on the principles of the present invention. Along these lines, each manufacturer could customize their BLAS subroutines in accordance with these principles.

It should also be recognized that other variations are possible, such as versions in which a higher level software module interfaces with existing linear algebra processing modules, such as a BLAS or other LAPACK module, to incorporate the principles of the present invention.

Moreover, the principles and methods of the present invention could be embodied as a computerized tool stored on a memory device, such as independent diskette 800, that contains a series of matrix subroutines to solve scientific and engineering problems using matrix processing, as modified by the technique described above. The modified matrix subroutines could be stored in memory as part of a math library, as is well known in the art. Alternatively, the computerized tool might contain a higher level software module to interact with existing linear algebra processing modules.

It should also be obvious to one of skill in the art that the instructions for the technique described herein can be downloaded through a network interface from a remote storage facility.

5 All of these various embodiments are intended as included in the present invention, since the present invention should be appropriately viewed as a method to enhance the computation of matrix subroutines, as based upon recognizing how linear algebra processing can be more efficient by using the principles of the present invention.

10 In yet another exemplary aspect of the present invention, it should also be apparent to one of skill in the art that the principles of the present invention can be used in yet another environment in which parties indirectly take advantage of the present invention.

15 For example, it is understood that an end user desiring a solution of a scientific or engineering problem may undertake to directly use a computerized linear algebra processing method that incorporates the method of the present invention. Alternatively, the end user might desire that a second party provide the end user the desired solution to the problem by providing the results of a computerized linear algebra processing method that incorporates the method of the present invention. These results might be provided to the end user by a
20 network transmission or even a hard copy printout of the results.

YOR920030169US1

The present invention is intended to cover all of these various methods of implementing and of using the present invention, including that of the end user who indirectly utilizes the present invention by receiving the results of matrix processing done in accordance with the principles herein.

5 In conclusion, many new architectures, for economic considerations, eliminate hardware instructions that might be desirable in scientific computations. Here, a case was encountered where the lack of an instruction would seem to imply that a performance penalty would be paid in the context of matrix multiplication, which is a pervasive algorithm and a workhorse subroutine for
10 such applications. The use of the pseudo-matrix concept allows one to overcome this disadvantage while incurring little extra work for the programmer and almost no performance penalty.

 While the invention has been described in terms of an exemplary embodiment, those skilled in the art will recognize that the invention can be
15 practiced with modification within the spirit and scope of the appended claims.

 Further, it is noted that, Applicants' intent is to encompass equivalents of all claim elements, even if amended later during prosecution.